MULTI-OBJECTIVE LLM ALIGNMENT USING MULTIVARIATE STOCHASTIC DOMINANCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Though Large Language Models (LLMs) are increasingly performant on a wide range of tasks, multi-objective alignment with a variety of evaluative signals is a key challenge. Existing methods typically optimize scalar preference signals, seek to balance competing objectives, or learn policies spanning different trade-offs, without explicitly targeting joint improvement over a reference policy. In this paper, we introduce **Stochastically Dominant Multi-Objective Alignment (SDMA)**, a framework for multi-dimensional LLM alignment without committing to particular trade-offs between evaluative dimensions. Instead, we employ first-order multivariate stochastic dominance to guide improved performance for all monotonic functions of these different dimensions. We demonstrate balanced improvement in our results for multiple dimensions, including subsets of helpfulness, harmlessness, humor, honesty, and instruction-following.

1 INTRODUCTION

Large Language Models (LLMs) have demonstrated strong capabilities across a wide range of tasks (Brown et al., 2020; Srivastava et al., 2022). As these models have become increasingly integrated into everyday life, better aligning the behavior of a reference policy with human preferences has become more important than ever before. Reinforcement Learning from Human Feedback (RLHF) has emerged as a prevalent approach to this problem (Christiano et al., 2017; Ziegler et al., 2019; Stiennon et al., 2020). However, standard RLHF typically represents preferences through a scalar reward. This overlooks the multifaceted nature of responses and their competing dimensions, such as helpfulness and harmlessness (Bai et al., 2022). Instead of focusing on a particular trade-off between these dimensions, supporting outputs across a range of trade-offs can be more useful (Yang et al., 2024; Zhou et al., 2024).

Multi-objective alignment methods for LLMs use multiple reward signals and often seek to approximate or parameterize a Pareto frontier (Rame et al., 2023; Yang et al., 2024; Zhou et al., 2024). However, Pareto optimality, though desirable, does not by itself provide performance improvement guarantees since many outputs along the frontier are less desirable than reference policy outputs for some utility functions. Distributional alignment methods such as Alignment via Optimal Transport (AOT) (Melnyk et al., 2024) address this by comparing complete scalar preference distributions. The univariate stochastic dominance that it seeks guarantees better expected utility than a reference policy for all monotonic transformations of an implicit single reward (Müller & Stoyan, 2002). However, AOT does not extend to multiple reward dimensions.

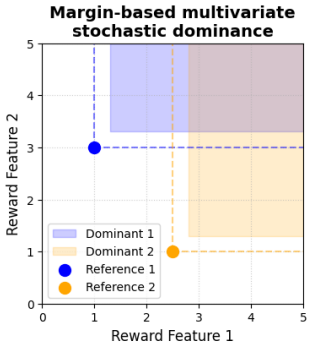


Figure 1: Two reference policy output samples evaluated under two reward features define Pareto dominance regions. Multivariate stochastic dominance requires an equal proportion of outputs from each of the two regions.

To address this gap, we develop **Stochastically Dominant Multi-Objective Alignment** (SDMA)[1], which seeks multivariate stochastic dominance (Armbruster & Luedtke, 2015; Rioux et al., 2024) (by a margin (Farajzadeh et al., 2026)). This ensures that the new policy has no less expected utility than the reference policy for all

[1]Code will be released upon acceptance.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

utility functions that are monotonic functions of the reward features (i.e., strictly greater reward vectors cannot have lower utility). Under this formulation, improvement on one reward dimension cannot compensate for degradation on another. To our knowledge, this is the first multi-objective LLM alignment framework to target broad monotonic utility guarantees using a single policy.

The main contributions of our paper are:

- Extending the expected utility guarantees of AOT (Melnyk et al., 2024) from the univariate to the multivariate setting for LLMs using multivariate stochastic dominance and explicit rewards;
- Augmenting multivariate stochastic dominance used for LLM comparison (Rioux et al., 2024) with a margin (Farajzadeh et al., 2026) to guide multi-objective policy improvement;
- Demonstrating the balanced improvement provided by SDMA using evaluations on the Anthropic HH-RLHF dataset (Bai et al., 2022) with two to five objectives (helpfulness, harmlessness, humor, honesty, and instruction-following); and
- Showing continued improvement of SDMA when combined with downstream policy optimization (best of N output selection).

Our results suggest that some apparent trade-offs in multi-objective alignment are method-dependent rather than unavoidable relative to the reference policy. We establish multivariate stochastic dominance as a guiding principle that provides balanced alignment over potentially competing objectives.

2 Background and Related Work

RLHF and multi-objective alignment. A common RLHF pipeline first performs supervised fine-tuning (SFT), then learns a scalar reward from pairwise human comparisons using a Bradley–Terry model (Bradley & Terry, 1952), and finally optimizes the policy against that reward using an algorithm such as proximal policy optimization (PPO) (Schulman et al., 2017) while penalizing deviation from the supervised policy (Christiano et al., 2017; Ziegler et al., 2019; Stiennon et al., 2020; Ouyang et al., 2022). Direct Preference Optimization (DPO) (Rafailov et al., 2023) provides an alternative that derives a pairwise objective from KL-regularized RLHF and learns directly from preferred and dispreferred responses without separately fitting a reward model or performing PPO. Despite their different optimization procedures, standard RLHF and DPO typically compress the criteria underlying human comparisons into a single preference signal.

Multi-objective reinforcement learning represents outcomes using reward vectors (Vamplew et al., 2018). A common approach selects a preference vector $\mathbf{w} \geq \mathbf{0}$ and maximizes the expected scalarized reward $\mathbf{w}^\top \mathbb{E}[\mathbf{R}]$. Varying $\mathbf{w}$ yields policies representing different compromises and can approximate a Pareto frontier. Multi-objective RLHF (MORLHF) (Zhou et al., 2024) applies this formulation to LLMs by optimizing weighted combinations of learned rewards. Safe RLHF instead maximizes expected helpfulness with an expected safety-cost constraint (Dai et al., 2024). MaxMin-RLHF addresses heterogeneity among annotator groups and maximizes the minimum expected utility across inferred group-specific reward functions (Chakraborty et al., 2024). Multi-objective DPO (MODPO) extends DPO using auxiliary reward differences as a preference-dependent margin and learns a policy for each selected preference vector (Zhou et al., 2024). These approaches make competing preferences explicit, but each produces a selected expected-reward compromise.

Pareto-oriented alignment. Other methods seek to represent a range of compromises more efficiently. Rewarded Soups fine-tunes one expert for each reward and constructs additional policies by interpolating their parameters (Rame et al., 2023). Rewards-in-Context conditions generation on target reward values and uses supervised fine-tuning to support preference adjustment at inference (Yang et al., 2024). Both aim to cover a Pareto frontier in expected reward space. Panacea (Zhong et al., 2024) embeds a supplied preference vector into the singular values of SVD-based LoRA layers and trains the resulting parameterization across the preference simplex. Its Pareto-front recovery result holds under stated representation and optimization assumptions. COS-DPO (Ren et al., 2025) similarly learns a weight-conditioned DPO model and uses objective weights to control its behavior after training. These methods maintain one learned parameterization, but the preference input induces a different effective policy for each weighting. They therefore represent a conditional family

of Pareto solutions. In contrast, we seek a policy that is not preference-conditioned but still improves upon the reference before any additional preference weighting is applied.

Complementary work addresses Pareto coverage and optimization conflict. HaM jointly trains several policies by maximizing their hypervolume (Mukherjee et al., 2024), while GAPO adaptively balances per-objective gradients (Li et al., 2025a). SIPO self-generates candidate Pareto-optimal responses to reduce conflicts in preference data (Li et al., 2025b), and Reward Consistency Sampling constructs comparisons supported across the considered reward objectives (Xu et al., 2026). These methods show that observed trade-offs can partly reflect the data or optimization procedure, but they do not formulate distributional improvement relative to a fixed reference.

Distributional alignment. Expected rewards do not describe how gains and losses are distributed across responses. First-order stochastic dominance compares complete scalar distributions by requiring one distribution to attain every reward threshold with at least as much probability as another (Ogryczak & Ruszczynski, 2002). Alignment via Optimal Transport (AOT) applies this criterion to scalar implicit preference statistics (Melnyk et al., 2024). Its unpaired formulation encourages the implicit rewards of preferred responses to dominate those of dispreferred responses. Its paired formulation compares the distributions of chosen-to-rejected log-probability ratios under the learned and reference policies. AOT therefore supplies a distributional objective, but for one scalar comparison statistic rather than multiple response-level rewards.

Distributional Pareto-Optimal MORL extends Pareto optimality from expected vector returns to multivariate return distributions (Cai et al., 2023). It learns policies that are not multivariate stochastically dominated by another attainable policy. UC-MOA transfers this notion to LLM personalization by conditioning a model on symbolic tokens associated with nonlinear monotone utility functions (Cheng et al., 2025). Its guarantee concerns the resulting utility-conditioned subpolicies assuming each one optimizes its associated expected utility. DPMORL and UC-MOA therefore seek members of a distributional Pareto set rather than a policy dominating a specified reference.

(Rioux et al., 2024) develop an optimal-transport statistic and hypothesis test for multivariate almost stochastic dominance and demonstrate it through multi-metric LLM benchmarking . Their contribution concerns model comparison rather than alignment. RAD is closer to our reference-relative setting (Chittepu et al., 2026). It replaces Safe RLHF's expected-cost constraint with stochastic dominance of the policy-induced safety-cost distribution relative to a reference. It retains an expected helpfulness objective and imposes distributional improvement on one safety signal rather than across multiple reward dimensions.

Pluralistic stochastic dominance (PSD). PSD requires stochastic dominance across a family of nonnegative combinations of reward features for imitation learning (Farajzadeh et al., 2026). It uses optimal transport to associate policy samples with demonstrations and penalizes reward shortfalls under the resulting assignment. We transfer this principle to multi-objective LLM alignment and compare an aligned policy with a fixed SFT reference under monotonic utility functions.

Our objective differs from both Pareto-front recovery and distributional Pareto optimality. Pareto-oriented methods produce a collection or conditioned family of preference-specific compromises. Distributional Pareto methods seek policies that are not dominated by another attainable policy. We instead seek one preference-independent policy that distributionally improves upon a specified reference across the entire space of monotonic utility functions.

3 Approach

Problem formulation: We consider a prompt distribution $x \sim \mathcal{D}$, a (stochastic) policy $\pi : \mathcal{X} \to \Delta_{\mathcal{Y}}$ that produces a distribution of outputs $y \in \mathcal{Y}$, and K reward models $\{r_i : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}\}$. Together, the reward models assign each prompt-response pair a vector of **reward features** assessing the output given the prompt:

$$\mathbf{r}(x, y) = [r_1(x, y) \quad \cdots \quad r_K(x, y)]^\top \in \mathbb{R}^K.$$

Given a reference policy π_0 (e.g., instantiated using supervised fine-tuning), our aim is to leverage the available reward features to produce an improved policy π.

As the **core motivating question** for our approach, we ask: *how can an LLM be improved while making the fewest additional assumptions beyond the available reward features?*

3.1 Stochastic Dominance for Multi-Objective LLM Alignment

To address our core question, we consider the most general functions for combining multiple objectives while still preserving order: *monotonic utility functions*.

Definition 1 *Utility function* $u : \mathbb{R}^K \to \mathbb{R}$*, is a* **monotonic function** *of the reward features:* $u \in \mathcal{U}_{mono} \iff (\mathbf{r}(x', y') \geq \mathbf{r}(x, y) \implies u(\mathbf{r}(x', y')) \geq u(\mathbf{r}(x, y)) \; \forall x, x', y, y')$.

This set of functions admits all utility functions that preserve an ordering between outputs defined by the multivariate reward functions *without* assuming any additional trade-offs between reward feature dimensions or the shape/smoothness of the utility function. For example, a set of reward features might each need to be high (a logical conjunction) to provide aggregate utility. Linear functions commonly adopted for balancing multiple objectives cannot represent this relationship. For a policy π, the corresponding random utility of responses is: $u(\mathbf{r}(X, \pi(X)))$, given $X \sim \mathcal{D}$. Multivariate first-order stochastic dominance (Def. 2) is defined by performance guarantees for this set of utility functions.

Definition 2 (Müller & Stoyan (2002)) *Random vector* $\mathbf{r}(X, \pi(X))$ *has* **multivariate stochastic dominance** *over* $\mathbf{r}(X, \pi_0(X))$*,*

$$\mathbf{r}(X, \pi(X)) \succeq_1 \mathbf{r}(X, \pi_0(X)) \iff \mathbb{E}_{\substack{X \sim \mathcal{D}; \\ Y|X \sim \pi}} [u(\mathbf{r}(X, Y))] \geq \mathbb{E}_{\substack{X \sim \mathcal{D}; \\ Y|X \sim \pi_0}} [u(\mathbf{r}(X, Y))] \; \forall u \in \mathcal{U}_{mono}. \quad (1)$$

In other words, for every monotonic utility function, the policy π must have expected utility no worse than the expected utility of the reference distribution π_0.

We note that in the univariate/scalar setting, first-order stochastic dominance is equivalently defined: (a) as providing no less expected utility for monotonic reward functions (Def. 2); and (b) distributionally by: $X \succeq_1 Y \iff \mathbb{P}(X \geq c) \geq \mathbb{P}(Y \geq c) \; \forall c \in \mathbb{R}$. However, in the multivariate setting, the distributional definition,

$$\mathbb{P}(\mathbf{X} \geq \mathbf{c}) \geq \mathbb{P}(\mathbf{Y} \geq \mathbf{c}) \; \forall \mathbf{c} \in \mathbb{R}^K, \quad (2)$$

is weaker than the utility-based Definition 2 (i.e., (1) $\implies$ (2), but (2) $\not\Rightarrow$ (1)) (Kopa & Petrová, 2018). We employ this stronger utility-based definition of multivariate stochastic dominance because: (a) it provides broader guarantees for performance than the distributional definition; and (b) it leads to more tractable detection of violations, as we describe next.

3.2 Matched Subdominance Minimization

Directly enforcing Equation (1) would require verifying stochastic dominance for every utility function $u \in \mathcal{U}_{\text{mono}}$. Unlike linear utility functions, this cannot be easily achieved using a set of (moment-based) constraints (Abbeel & Ng, 2004; Syed et al., 2008). Following Armbruster & Luedtke (2015), we instead use a definition based on a matching between samples generated by the learned and reference policies.

Theorem 3 (Armbruster & Luedtke (2015)) $R \succeq_1 R^0$ *for two uniform sample distributions of matching size if and only if a one-to-one pairing between* $\mathbf{r}_i$ $(i \in [N])$ *and* $\mathbf{r}^0_j$ $(j \in [N])$ *exists such that* $\mathbf{r}_i \geq \mathbf{r}^0_j$.

Pareto dominance is trivially achieved by $\pi = \pi_0$. To avoid this solution, we follow Farajzadeh et al. (2026) and seek paired Pareto dominance by a margin based on hyperparameters α and β. For a policy sample i and a reference sample j, the subdominance cost (Ziebart et al., 2022) is:

$$\text{subdom}_{\boldsymbol{\alpha},\beta}(\mathbf{r}_i, \mathbf{r}^0_j) = \sum_{k=1}^{K} \left[\alpha_k \left(r^0_{j,k} - r_{i,k}\right) + \beta\right]_+, \quad (3)$$

where $[z]_+ = \max\{z, 0\}$, $\boldsymbol{\alpha} \in \mathbb{R}^K_{>0}$, and $\beta \geq 0$. For $\beta = 0$, this is ordinary Pareto dominance. A positive β requires reward improvement by a margin. The weights $\boldsymbol{\alpha}$ control the relative subdominance penalties and margins during optimization and are distinct from the user preference vector $\mathbf{w}$ in Equation (1).

We next employ an optimal transport extension of Theorem 3 (Rioux et al., 2024) allowing partial/probabilistic matching. Let p_i and p_j^0 denote the probability masses assigned to policy sample y_i and reference sample y_j^0. Matched subdominance between the two policies is defined as:

$$\mathrm{OT}_{\mathrm{subdom}_{\boldsymbol{\alpha},\beta}}(\pi, \pi_0) = \min_{\boldsymbol{\gamma} \geq 0} \sum_{i,j} \gamma_{ij} \, \mathrm{subdom}_{\boldsymbol{\alpha},\beta}(\mathbf{r}_i, \mathbf{r}_j^0) \quad \text{subject to} \quad \begin{cases} \sum_j \gamma_{ij} = p_i, & \forall i, \\ \sum_i \gamma_{ij} = p_j^0, & \forall j. \end{cases} \tag{4}$$

Here, γ_{ij} is the probability mass assigned to matching policy sample y_i with reference sample y_j^0. The marginal constraints ensure $\boldsymbol{\gamma}$ is a valid coupling of the samples generated by the two policies.

Since every term in Equation (4) is nonnegative, an objective value of zero means that positive mass is assigned only to pairs with zero subdominance. This is a sufficient condition for multivariate stochastic dominance (Theorem 4).

Theorem 4 (Farajzadeh et al. (2026)) *An optimal transport objective of* 0 *implies first-order multivariate stochastic dominance:*

$$\mathrm{OT}_{\mathrm{subdom}_{\boldsymbol{\alpha},\beta}}(\pi, \pi_0) = 0 \quad \Longrightarrow \quad \mathbf{r}(X, \pi(X)) \succeq_1 \mathbf{r}(X, \pi_0(X)). \tag{5}$$

3.3 SDMA Policy Optimization

Given output samples from policy π_ϕ, and a subdominance-minimizing matching solution ($\gamma^\star$) from (4), we construct a learning signal for each policy sample using a weighted-best-match construction of Farajzadeh et al. (2026), as shown in Algorithm 1.

Algorithm 1 Policy improvement via Stochastically Dominant Multi-Objective Alignment (SDMA)

Input: Policy π_ϕ, matching probabilities from optimal transport γ^*_{ij}, sets of policy reward vectors, $\{\mathbf{r}_j^0\}$ and $\{\mathbf{r}_i\}$ sampled from the reference policy π^0 and the (improved) policy π_ϕ and hyperparameters $\boldsymbol{\alpha}$, β.

Output: An improved policy π_ϕ

1: Compute $c_{i,j} = \mathrm{subdom}_{\boldsymbol{\alpha},\beta}(\mathbf{r}_i, \mathbf{r}_j^0)$ for all pairs i, j
2: Compute $\bar{c}_j = \sum_{i'} \gamma^\star_{i'j} c_{i'j}$, and $i_j^\star \in \mathrm{argmin}_{i'} \, c_{i'j}$
3: Compute $a_i = \sum_j \left(\gamma^\star_{ij} c_{ij} - \mathbb{I}[i = i_j^\star] \, \bar{c}_j\right)$
4: Apply update $\phi \leftarrow \phi - \eta \sum_i a_i \nabla_\phi \log \pi_\phi(y_i \mid x_i)$

Using the average cost for each reference sample (Line 2), we define a learning weight a_i for each policy sample in Line 3. The first term of this weight assigns each policy sample the subdominance cost of its transported mass. The second credits it when it is the lowest-cost match for a reference sample. Together, these center the contribution of each reference sample while favoring its best available policy match. Finally, we apply policy gradient optimization step in Line 4. This increases the probability of responses favored by weighted best matching and decreases the probability of responses with larger subdominance.

The subdominance weights $\boldsymbol{\alpha}$ may be held fixed or optimized jointly with the policy. When learned, we update $\boldsymbol{\alpha}$ by ascending the matched-subdominance objective while holding $\boldsymbol{\gamma}^\star$ fixed, and then project it back onto the constrained simplex. This places greater emphasis on objectives with unresolved subdominance violations.

4 Experiments

We experimentally investigate two key questions of our multivariate stochastic dominance approach:

- Q1. Does the broad requirement of guaranteeing performance for all monotonic functions of competing reward signals allow LLM improvement in practice?
- Q2. Does our method produce more balanced improvement across reward signals than existing multi-objective optimization methods for LLMs?

4.1 Experimental Setup

We conduct our experiments using the Anthropic HH-RLHF dataset (Bai et al., 2022). All methods are initialized from the same reference policy, π_{SFT}, obtained by supervised fine-tuning Llama-2-7B (Touvron et al., 2023) on the preferred responses in the training data. While training SDMA, we iterate over the training prompts in shuffled mini-batches of size 30. For each mini-batch, we generate three responses per prompt using the current policy and compute the optimal transport matching between these responses and the prompts' base responses, which are sampled once from π_{SFT} before training. We start our experiments with a two-objective setting consisting of helpfulness and harmlessness. Inspired by the three-objective evaluation in RiC (Yang et al., 2024), we then add humor as a third objective to study alignment under a broader range of potentially competing reward dimensions. We subsequently add honesty as a fourth objective and instruction-following as a fifth objective. Each response is evaluated separately on every objective, producing reward vectors ranging from two to five dimensions. We use publicly available, off-the-shelf reward models for our objectives. Our helpfulness (Ray2333, 2023b) and harmlessness (Ray2333, 2023a) reward models are GPT-2-large models trained on the corresponding subsets of HH-RLHF and are the same reward models used for helpfulness and harmlessness in RiC (Yang et al., 2024) . For humor, we use a RoBERTa-based humor classifier (Baranov et al., 2023), and for honesty and instruction following, the corresponding objective outputs of ArmoRM (Wang et al., 2024).

4.2 Baselines

We compare our method with a broad set of multi-objective and distributional alignment approaches introduced in §2: MORLHF (Zhou et al., 2024) optimizes a uniformly weighted combination of the reward dimensions; MaxMin-RLHF (Chakraborty et al., 2024) optimizes the lowest-scoring objective; DPO (Rafailov et al., 2023) is a standard preference-optimization baseline; MODPO is its multivariate extension (Zhou et al., 2024). We further compare against methods designed to represent different compromises among reward dimensions. Rewarded Soups (Rame et al., 2023) interpolates the parameters of experts trained separately for each objective, whereas Reward in Context (RiC) (Yang et al., 2024) conditions response generation on target reward values. For Rewarded Soups, we report the equal-weight interpolation of the objective-specific experts, and for RiC, the main tables report the all-max setting, which specifies the maximum target reward for every objective. As the closest distributional alignment baseline, we include AOT (Melnyk et al., 2024), which uses optimal transport to impose stochastic dominance on a scalar preference statistic.

In the two-objective setting, the preference-based baselines are trained on the human preference comparisons in HH-RLHF. Because the dataset does not contain humor, honesty, or instruction-following comparisons, for $K \geq 3$ we construct them: for each additional objective, we sample 16 responses per prompt from π_{SFT} and pair the highest- and lowest-scoring responses under that objective's reward model. These pairs supplement the human helpfulness and harmlessness comparisons.

Finally, we apply the same best-of-16 selection procedure to candidates generated by both (π_{SFT}) and SDMA. For each prompt, the procedure generates 16 candidate responses and selects the one with the highest uniformly weighted sum of the objective rewards. Best-of-16 on (π_{SFT}) measures what inference-time selection can achieve from the reference distribution, while applying it to SDMA tests whether SDMA provides a stronger base for the same downstream procedure. Because selection uses the evaluation reward models at inference time, we report both as oracle analyses rather than directly comparable trained baselines.

4.3 Evaluation Protocol and Metrics

Existing evaluations of preference alignment commonly emphasize mean rewards and win rates, while multi-objective methods additionally compare Pareto fronts across objective weightings (Rafailov et al., 2023; Rame et al., 2023; Yang et al., 2024). These measures characterize average performance and reward trade-offs, but they do not establish whether improvement extends to the broader families of utility functions. To instead assess distributional properties, we evaluate each method on 1,000 held-out prompts using the same fixed set of reference responses generated by π_{SFT}. We conduct five independent generation trials for each method and report the mean of each metric across trials. To calibrate the results against variation from stochastic generation, we

also include a base-versus-base control comparing fresh responses from π_{SFT} with those fixed reference responses. To capture different aspects of broad improvement over the reference policy, we use the following complementary metrics:

Matched Pareto Dominance (MPD). Using responses sampled from the evaluated and reference LLM policies, we find a one-to-one matching $\sigma^\star \in \arg\min_{\sigma\in\mathfrak{S}_N} \sum_{i=1}^{N} \mathrm{subdom}_{\mathbf{1},0}(\mathbf{r}_i, \mathbf{r}^0_{\sigma(i)})$ (where $\mathfrak{S}_N$ denotes permutations of N samples) that minimizes the total subdominance (Ziebart et al., 2022) cost using the Hungarian algorithm (Kuhn, 1955). We then report the percentage of matched pairs for which the evaluated response Pareto-dominates base: $\mathrm{MPD} = \frac{100}{N}\sum_{i=1}^{N} \mathbb{I}[\mathbf{r}_i \geq \mathbf{r}^0_{\sigma^\star(i)}]$, where $\geq$ denotes component-wise comparison. This measures how close the policy is to guaranteeing no worse performance for all monotonic utility functions.

Sampled Linear Dominance. We quantify the extent of distributional improvement over the reference using a score adapted from almost stochastic dominance (Leshno & Levy, 2002). The score measures the share of the total absolute survival-curve gap favoring the evaluated policy after strengthening the reference by a margin (m), averaged across 1,000 preference vectors sampled uniformly from the simplex that are used to define a linear utility function.
Let $S_{\pi,\mathbf{w}}(c)$ denote the empirical probability that the scalarized reward exceeds threshold c, and let $S^m_{\pi_0,\mathbf{w}}(c)$ denote the corresponding probability after shifting the reference rewards upward by m times their reference range, i.e., by $mR_{\mathbf{w}}$ with $R_{\mathbf{w}} = \max_j \mathbf{w}^\top \mathbf{r}^0_j - \min_j \mathbf{w}^\top \mathbf{r}^0_j$, where $\mathbf{r}^0_j$ is the reward vector of the j-th reference response and $R_{\mathbf{w}}$ is computed separately for each preference vector $\mathbf{w}$. Equivalently, $S^m_{\pi_0,\mathbf{w}}(c) = S_{\pi_0,\mathbf{w}}(c - mR_{\mathbf{w}})$. We define $\epsilon_m(\mathbf{w}) = \sum_{c\in\mathcal{C}}[S^m_{\pi_0,\mathbf{w}}(c) - S_{\pi,\mathbf{w}}(c)]_+ / \sum_{c\in\mathcal{C}} |S^m_{\pi_0,\mathbf{w}}(c) - S_{\pi,\mathbf{w}}(c)|$, and $Lin_m = \frac{100}{|\mathcal{W}|}\sum_{\mathbf{w}\in\mathcal{W}}(1-\epsilon_m(\mathbf{w}))$. Here $\mathcal{W}$ and $\mathcal{C}$ are the evaluated preference vectors and thresholds, respectively.

Worst-Group Sampled Linear Dominance. To detect weaknesses obscured by averaging across all random preference vectors, we also report the lowest group-average score. Each group $\mathcal{W}_k$ assigns weight 0.8 to objective k and samples the remaining 0.2 uniformly from the other objectives' simplex: $Lin^{\mathrm{WGD}}_m = \min_{k\in\{1,\ldots,K\}} \frac{100}{|\mathcal{W}_k|}\sum_{\mathbf{w}\in\mathcal{W}_k}\big(1-\epsilon_m(\mathbf{w})\big)$. For $K=2$, each group contains a single preference vector. We compute the minimum separately within each generation trial before averaging across trials.

Per-objective effect sizes. We report Cohen's d (Cohen, 2013) separately for each reward dimension, $d_k = (\mu_k - \mu^0_k)/\sqrt{\frac{1}{2}(v_k + v^0_k)}$, where μ_k, v_k and μ^0_k, v^0_k are the empirical reward means and variances for the evaluated and reference policies, respectively. We compute d_k independently for each generation trial and report its mean across five trials.

4.4 Main Results

Five-objective alignment evaluation. Table 1 reports results for all five reward signals. SDMA achieves 94.1% on MPD metric compared with 22.1% for the strongest non-oracle baseline, MORLHF, indicating much better performance across monotonic utility functions. SDMA has pos-

Table 1: Results with five objectives: helpfulness, harmlessness, humor, honesty, and instruction-following. The first five metrics are percentages. Effect sizes are Cohen's d relative to the reference policy. Best-of-16 rows are oracle references that use the evaluation reward models at inference time. Top values for non-oracle methods are shown in bold.

Method	MPD	$\mathrm{Lin}_{.05}$	$\mathrm{Lin}_{.10}$	$\mathrm{Lin}^{\mathrm{WG}}_{.05}$	$\mathrm{Lin}^{\mathrm{WG}}_{.10}$	d_{help}	d_{harm}	d_{humor}	d_{honesty}	d_{instr}
SDMA	**94.1**	**99.9**	**98.8**	**99.4**	**44.3**	**+0.36**	**+0.50**	+0.69	**+0.88**	**+1.03**
Rewarded Soups	9.2	1.4	0.0	0.0	0.0	+0.10	−0.05	+0.04	+0.24	+0.24
MODPO	4.8	10.8	0.0	0.0	0.0	+0.28	−0.22	−0.01	−0.05	+0.33
DPO	19.4	97.1	82.3	2.5	0.0	−0.11	+0.30	**+1.42**	+0.67	+0.79
MaxMin-RLHF	6.1	19.3	0.7	1.6	0.0	+0.02	+0.07	+0.32	+0.09	−0.05
MORLHF	22.1	80.6	18.5	1.8	0.0	+0.05	+0.16	+0.50	+0.38	+0.18
AOT	6.8	80.1	53.4	0.1	0.0	−0.24	+0.26	+1.25	+0.25	+0.12
RiC	1.9	5.3	0.2	0.0	0.0	+0.33	−0.37	+0.05	−0.11	+0.20
Reference-policy	4.2	0.0	0.0	0.0	0.0	−0.01	+0.00	−0.01	+0.01	+0.00
BoN-16 (π_{SFT}, oracle)	29.6	92.1	70.4	5.4	0.2	+0.34	+0.00	+0.50	+1.03	+0.89
BoN-16 (SDMA, oracle)	95.5	100.0	99.1	95.9	27.3	+0.88	+0.39	+0.81	+1.76	+1.92

itive effect sizes on all dimensions: (helpfulness $d = +0.36$, harmlessness $d = +0.50$, humor $d = +0.69$, honesty $d = +0.88$, and instruction-following $d = +1.03$), while its sampled linear dominance scores are 99.9% at $m = 0.05$ and 98.8% at $m = 0.10$. The worst-group linear dominance scores remain high for $m = 0.05$ and decrease significantly for $m = 0.10$, but remain much better than all other methods. Additionally, SDMA's worst and average effect sizes are larger than all other methods. Together, these results show that SDMA improves both individual objectives and the broader reward distributions rather than producing a gain visible only in their means.

Baseline methods all produce worse performance than SDMA on all dominance-based metrics. This indicates more significant suboptimality within the class of monotonic utility functions and less dominant performance on linear utility functions. Further, though some baseline methods perform better on one reward signal (AOT and DPO on humor), they also exhibit performance degradations in at least one reward signal. This shows that these methods are generally not suitable for supporting all preference-agnostic downstream tasks and support an affirmative answer for Q2.

Lastly, applying the best-of-16 oracle to candidates from π_{SFT} and SDMA in the last two rows of Table 1 tests whether SDMA provides a stronger starting point for subsequent downstream reward-guided improvement methods. In this five-objective setting, matched dominance rises from 29.6% with π_{SFT} candidates to 95.5% with SDMA candidates. We note significantly better performance of SDMA-based BoN versus SFT-based BoN across all metrics evaluated. These large differences show that SDMA produces jointly high performant responses, enabling the same downstream procedure to achieve substantially better results.

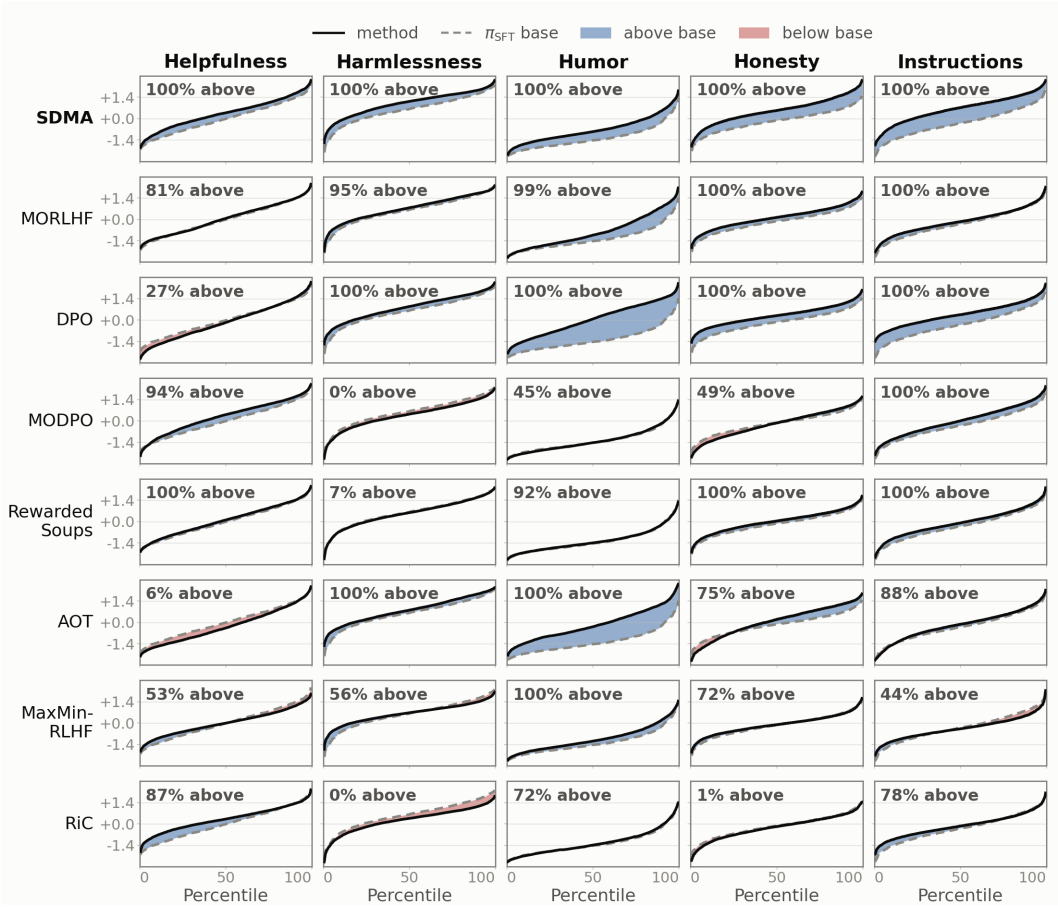


Figure 2: Per-objective reward quantile functions in the five-objective setting based on 5000 responses (1000 for SFT base) from 1000 prompts. Percentages indicate the fraction of plotted percentiles at which each method is at least as high as the reference. The y-axes are in standard deviations of the reference policy. Only SDMA is consistently above the base for each objective.

Figure 2 provides a univariate distributional perspective to the improvement that each method makes to the reference LLM. Specifically, it shows the reward attained at each percentile of the response distribution. SDMA remains at or above the reference at every plotted percentile on all five objectives, with substantial separation across much of each curve. Its gains therefore extend throughout the reward distributions rather than being limited to improvements in their means, while every trained baseline falls below the reference on at least one objective. This is not by coincidence; univariate stochastic dominance in each dimension (and joint distributional stochastic dominance) is implied by our stronger utility-based multivariate stochastic dominance (Kopa & Petrová, 2018).

Alignment evaluation across different numbers of objectives. We next report our alignment evaluations with other numbers of reward objectives. From $K = 2$ to $K = 4$, we consider the first K reward signals from: *helpfulness*, *harmlessness*, *humor*, and *honesty*. Though full results of this comparison are presented in the appendix, we summarize the performance of SDMA and the best non-oracle baseline method in Table 2. SDMA exhibits better performance across all dominance-based metrics for each choice of K. Additionally, while SDMA provides balanced improvement across each of the K objectives, other baselines sometimes perform better on one or some of the dimensions at the expense of others. This is particularly evident for $K = 3$ in which SDMA is outperformed on d_{help}, d_{harm}, and d_{humor} by a different baseline for each dimension. These results further support affirmative answers for Q1 and Q2.

Table 2: **SDMA against the best canonical baseline, $K = 2$ to 5.** Each baseline value is the maximum over the seven canonical baselines, taken independently per column, with the method attaining it below; "all" marks columns where every baseline scores 0.0. d_k is the per-objective effect size against the reference policy. Bold marks the better of each pair.

K	Method	MPD	$\text{Lin}_{.05}$	$\text{Lin}_{.10}$	$\text{Lin}^{\text{WG}}_{.05}$	$\text{Lin}^{\text{WG}}_{.10}$	d_{help}	d_{harm}	d_{humor}	d_{honest}	d_{instr}
2	SDMA	**99.9**	**100.0**	**98.2**	**100.0**	**99.9**	+0.58	**+1.20**	–	–	–
	Best baseline	90.6	66.9	16.1	44.9	2.1	**+0.69**	+0.53	–	–	–
		DPO	DPO	MORLHF	AOT	DPO	RiC	MORLHF			
3	SDMA	**98.2**	**99.9**	**97.9**	**99.0**	**79.6**	+0.57	+0.66	+0.99	–	–
	Best baseline	21.2	86.4	76.6	4.2	0.0	**+0.75**	**+0.69**	**+1.78**	–	–
		Soups	MORLHF	MORLHF	MaxMin	all	RiC	MORLHF	AOT		
4	SDMA	**98.1**	**100.0**	**98.5**	**99.9**	**76.1**	+0.33	**+0.72**	+1.32	+0.64	–
	Best baseline	19.8	93.7	73.9	4.7	0.0	**+0.72**	+0.45	**+2.47**	**+0.84**	–
		DPO	DPO	DPO	MaxMin	all	RiC	MORLHF	AOT	DPO	
5	SDMA	**94.1**	**99.9**	**98.8**	**99.4**	**44.3**	**+0.36**	**+0.50**	+0.69	**+0.88**	**+1.03**
	Best baseline	22.1	97.1	82.3	2.5	0.0	+0.33	+0.30	**+1.42**	+0.67	+0.79
		MORLHF	DPO	DPO	DPO	all	RiC	DPO	DPO	DPO	DPO

5 Conclusions and Limitations

In this paper, we sought to improve Large Language Models given multiple reward signals while making minimal additional assumptions about the utility of outputs for end uses. We employed a utility-based view of multivariate stochastic dominance to formulate this aim. Leveraging recent work using optimal transport formulations to compare LLMs (Rioux et al., 2024) and improve imitation learning policies (Farajzadeh et al., 2026), we applied our approach to optimize LLMs using multiple competing reward signals. Our experimental results show that, though imposing a stricter notion of improvement, our approach succeeds in improving policies evaluated across two to five reward features. Compared to existing LLM optimization methods, we demonstrate that improvements from our approach are more balanced across reward signal dimensions and extend over the entire distribution of outputs rather than simply in expectation.

The effectiveness of our approach depends on reliable reward signals. Positive biases can shift the policy of the resulting LLM in a manner that does not translate into greater monotonic utility. One approach to avoid this is to incorporate multiple redundant reward signals from different models. This avoids overfitting to the biases of any one of them. Experiments for scaling to these larger sets of correlated reward signals remain as important future work.

REFERENCES

Pieter Abbeel and Andrew Y. Ng. Apprenticeship learning via inverse reinforcement learning. In *Proc. International Conference on Machine Learning*, pp. 1–8, 2004.

Benjamin Armbruster and James Luedtke. Models and formulations for multivariate dominance-constrained stochastic programs. *IIE Transactions*, 47(1):1–14, 2015.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

Alexander Baranov, Vladimir Kniazhevsky, and Pavel Braslavski. You told me that joke twice: A systematic investigation of transferability and robustness of humor detection models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 13701–13715, 2023.

Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Xin-Qiang Cai, Pushi Zhang, Li Zhao, Jiang Bian, Masashi Sugiyama, and Ashley Llorens. Distributional pareto-optimal multi-objective reinforcement learning. *Advances in Neural Information Processing Systems*, 36:15593–15613, 2023.

Souradip Chakraborty, Jiahao Qiu, Hui Yuan, Alec Koppel, Furong Huang, Dinesh Manocha, Amrit Singh Bedi, and Mengdi Wang. Maxmin-rlhf: Alignment with diverse human preferences. *arXiv preprint arXiv:2402.08925*, 2024.

Zelei Cheng, Xin-Qiang Cai, Yuting Tang, Pushi Zhang, Boming Yang, Masashi Sugiyama, and Xinyu Xing. Uc-moa: Utility-conditioned multi-objective alignment for distributional pareto-optimality. *arXiv preprint arXiv:2503.10669*, 2025.

Yaswanth Chittepu, Ativ Joshi, Rajarshi Bhattacharjee, and Scott Niekum. Safe rlhf beyond expectation: Stochastic dominance for universal spectral risk control. *arXiv preprint arXiv:2603.10938*, 2026.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Jacob Cohen. *Statistical power analysis for the behavioral sciences*. routledge, 2013.

Juntao Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. Safe rlhf: Safe reinforcement learning from human feedback. In *International Conference on Learning Representations*, volume 2024, pp. 50750–50777, 2024.

Ali Farajzadeh, Danyal Saeed, Syed M Abbas, Rushit Shah, Aadirupa Saha, and Brian Ziebart. Imitation beyond expectation using pluralistic stochastic dominance. *Advances in Neural Information Processing Systems*, 38:27317–27344, 2026.

Milos Kopa and Barbora Petrová. Strong and weak multivariate first-order stochastic dominance. *Available at SSRN 3144058*, 2018.

Harold W Kuhn. The Hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

Moshe Leshno and Haim Levy. Preferred by "all" and preferred by "most" decision makers: Almost stochastic dominance. *Management Science*, 48(8):1074–1085, 2002.

Chengao Li, Hanyu Zhang, Yunkun Xu, Hongyan Xue, Xiang Ao, and Qing He. Gradient-adaptive policy optimization: Towards multi-objective alignment of large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11214–11232, 2025a.

Moxin Li, Yuantao Zhang, Wenjie Wang, Wentao Shi, Zhuo Liu, Fuli Feng, and Tat-Seng Chua. Self-improvement towards pareto optimality: Mitigating preference conflicts in multi-objective alignment. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 11010–11031, 2025b.

Igor Melnyk, Youssef Mroueh, Brian Belgodere, Mattia Rigotti, Apoorva Nitsure, Mikhail Yurochkin, Kristjan Greenewald, Jiri Navratil, and Jarret Ross. Distributional preference alignment of llms via optimal transport. *Advances in Neural Information Processing Systems*, 37: 104412–104442, 2024.

Subhojyoti Mukherjee, Anusha Lalitha, Sailik Sengupta, Aniket Deshmukh, and Branislav Kveton. Multi-objective alignment of large language models through hypervolume maximization. *arXiv preprint arXiv:2412.05469*, 2024.

Alfred Müller and Dietrich Stoyan. *Comparison methods for stochastic models and risks*. Wiley, 2002.

WLodzimierz Ogryczak and Andrzej Ruszczynski. Dual stochastic dominance and related mean-risk models. *SIAM Journal on Optimization*, 13(1):60–78, 2002.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744, 2022.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

Alexandre Rame, Guillaume Couairon, Corentin Dancette, Jean-Baptiste Gaya, Mustafa Shukor, Laure Soulier, and Matthieu Cord. Rewarded soups: towards pareto-optimal alignment by interpolating weights fine-tuned on diverse rewards. *Advances in Neural Information Processing Systems*, 36:71095–71134, 2023.

Ray2333. gpt2-large-harmless-reward_model. `https://huggingface.co/Ray2333/gpt2-large-harmless-reward_model`, 2023a. Hugging Face model repository, accessed August 11, 2026.

Ray2333. gpt2-large-helpful-reward_model. `https://huggingface.co/Ray2333/gpt2-large-helpful-reward_model`, 2023b. Hugging Face model repository, accessed August 11, 2026.

Yinuo Ren, Tesi Xiao, Michael Shavlovsky, Lexing Ying, and Holakou Rahmanian. Cos-dpo: Conditioned one-shot multi-objective fine-tuning framework. In Silvia Chiappa and Sara Magliacane (eds.), *Proceedings of the Forty-first Conference on Uncertainty in Artificial Intelligence*, volume 286 of *Proceedings of Machine Learning Research*, pp. 3525–3551. PMLR, 21–25 Jul 2025. URL `https://proceedings.mlr.press/v286/ren25a.html`.

Gabriel Rioux, Apoorva Nitsure, Mattia Rigotti, Kristjan Greenewald, and Youssef Mroueh. Multivariate stochastic dominance via optimal transport and applications to models benchmarking. *Advances in Neural Information Processing Systems*, 37:39190–39223, 2024.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*, 2022.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in neural information processing systems*, 33:3008–3021, 2020.

Umar Syed, Michael Bowling, and Robert E Schapire. Apprenticeship learning using linear programming. In *International Conference on Machine Learning*, pp. 1032–1039, 2008.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Peter Vamplew, Richard Dazeley, Cameron Foale, Sally Firmin, and Jane Mummery. Human-aligned artificial intelligence is a multiobjective problem. *Ethics and information technology*, 20(1):27–40, 2018.

Haoxiang Wang, Wei Xiong, Tengyang Xie, Han Zhao, and Tong Zhang. Interpretable preferences via multi-objective reward modeling and mixture-of-experts. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 10582–10592, 2024.

Zhihao Xu, Yongqi Tong, Xin Zhang, Jun Zhou, and Xiting Wang. Understanding conflicts in multi-objective alignment through reward consistency. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 5450–5472, 2026.

Rui Yang, Xiaoman Pan, Feng Luo, Shuang Qiu, Han Zhong, Dong Yu, and Jianshu Chen. Rewards-in-context: Multi-objective alignment of foundation models with dynamic preference adjustment. *arXiv preprint arXiv:2402.10207*, 2024.

Yifan Zhong, Chengdong Ma, Xiaoyuan Zhang, Ziran Yang, Haojun Chen, Qingfu Zhang, Siyuan Qi, and Yaodong Yang. Panacea: Pareto alignment via preference adaptation for llms. *Advances in Neural Information Processing Systems*, 37:75522–75558, 2024.

Zhanhui Zhou, Jie Liu, Jing Shao, Xiangyu Yue, Chao Yang, Wanli Ouyang, and Yu Qiao. Beyond one-preference-fits-all alignment: Multi-objective direct preference optimization. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 10586–10613, 2024.

Brian Ziebart, Sanjiban Choudhury, Xinyan Yan, and Paul Vernaza. Towards uniformly superhuman autonomy via subdominance minimization. In *International Conference on Machine Learning*, pp. 27654–27670. PMLR, 2022.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.

A USE OF LLMS

We used Large Language Models to assist with writing code, editing the manuscript for clarity and readability, and finding related work. The authors reviewed the generated code and manuscript edits and take responsibility for the final content of this work.

B RESULTS WITH TWO TO FOUR OBJECTIVES

Tables 3, 4, and 5 report all metrics for every method with two, three, and four objectives, in the same format as Table 1. Figures 3, 4, and 5 show the corresponding per-objective reward quantile functions. As with five objectives, SDMA is the only method that is consistently above the reference on every objective, whereas each baseline falls below the reference on at least one objective.

Table 3: Results with two objectives: helpfulness and harmlessness. The first five metrics are percentages. Effect sizes are Cohen's d relative to the reference policy. Best-of-16 rows are oracle references that use the evaluation reward models at inference time. Top values for non-oracle methods are shown in bold. For $K = 2$, each $\mathrm{Lin}^{\mathrm{WG}}$ group is a single weight vector, since fixing one weight at 0.8 fixes the other at 0.2.

Method	MPD	$\mathrm{Lin}_{.05}$	$\mathrm{Lin}_{.10}$	$\mathrm{Lin}^{\mathrm{WG}}_{.05}$	$\mathrm{Lin}^{\mathrm{WG}}_{.10}$	d_{help}	d_{harm}
SDMA	**99.9**	**100.0**	**98.2**	**100.0**	**99.9**	$+0.58$	$\mathbf{+1.20}$
Rewarded Soups	58.1	4.9	0.0	2.1	0.0	$+0.16$	$+0.04$
MODPO	3.9	35.5	0.8	0.0	0.0	$+0.39$	-0.10
DPO	90.6	66.9	7.1	36.3	2.1	$+0.26$	$+0.19$
MaxMin-RLHF	10.7	19.4	2.8	14.9	0.0	$+0.08$	$+0.14$
MORLHF	1.2	59.6	16.1	0.1	0.0	-0.20	$+0.53$
AOT	66.1	57.4	7.5	44.9	0.0	$+0.09$	$+0.29$
RiC	1.0	28.3	12.9	0.0	0.0	$\mathbf{+0.69}$	-0.57
Reference-policy	20.2	0.0	0.0	0.0	0.0	-0.01	$+0.00$
BoN-16 (π_{SFT}, oracle)	99.5	95.0	50.0	89.3	24.7	$+0.35$	$+0.39$
BoN-16 (SDMA, oracle)	99.9	100.0	99.4	100.0	100.0	$+0.89$	$+1.77$

Table 4: Results with three objectives: helpfulness, harmlessness, and humor. The first five metrics are percentages. Effect sizes are Cohen's d relative to the reference policy. Best-of-16 rows are oracle references that use the evaluation reward models at inference time. Top values for non-oracle methods are shown in bold.

Method	MPD	$\mathrm{Lin}_{.05}$	$\mathrm{Lin}_{.10}$	$\mathrm{Lin}^{\mathrm{WG}}_{.05}$	$\mathrm{Lin}^{\mathrm{WG}}_{.10}$	d_{help}	d_{harm}	d_{humor}
SDMA	**98.2**	**99.9**	**97.9**	**99.0**	**79.6**	$+0.57$	$+0.66$	$+0.99$
Rewarded Soups	21.2	7.3	0.0	0.0	0.0	$+0.04$	-0.01	$+0.28$
MODPO	1.5	42.3	0.4	0.0	0.0	-0.40	$+0.33$	$+0.34$
DPO	2.6	69.7	21.9	0.0	0.0	-0.34	$+0.41$	$+0.68$
MaxMin-RLHF	5.6	50.2	4.1	4.2	0.0	-0.09	$+0.11$	$+0.51$
MORLHF	1.9	86.4	76.6	2.1	0.0	-0.58	$\mathbf{+0.69}$	$+1.55$
AOT	0.4	77.9	65.4	0.0	0.0	-0.68	$+0.60$	$\mathbf{+1.78}$
RiC	0.5	17.3	6.6	0.0	0.0	$\mathbf{+0.75}$	-0.67	-0.19
Reference-policy	11.6	0.0	0.0	0.0	0.0	-0.01	$+0.00$	-0.01
BoN-16 (π_{SFT}, oracle)	10.3	87.6	65.5	16.2	0.5	-0.10	$+0.28$	$+1.30$
BoN-16 (SDMA, oracle)	99.9	100.0	99.6	100.0	99.0	$+0.45$	$+1.10$	$+1.47$

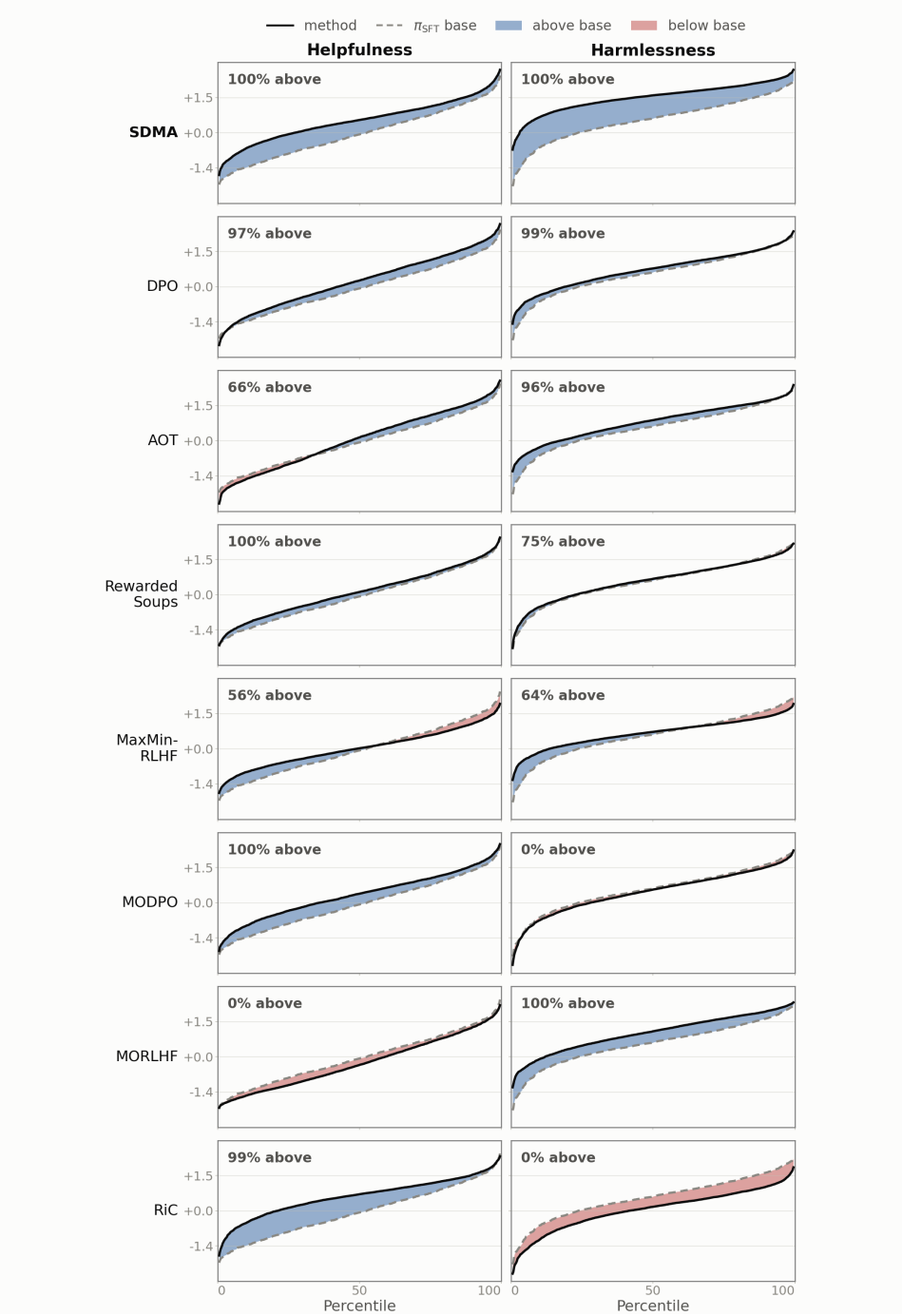


Figure 3: Per-objective reward quantile functions in the two-objective setting based on 5000 responses (1000 for SFT base) from 1000 prompts. Percentages indicate the fraction of plotted percentiles at which each method is at least as high as the reference. The y-axes are in standard deviations of the reference policy. Only SDMA is consistently above the base for each objective; DPO comes closest but falls up to 0.26 standard deviations below the base in the bottom 3.5% of helpfulness.

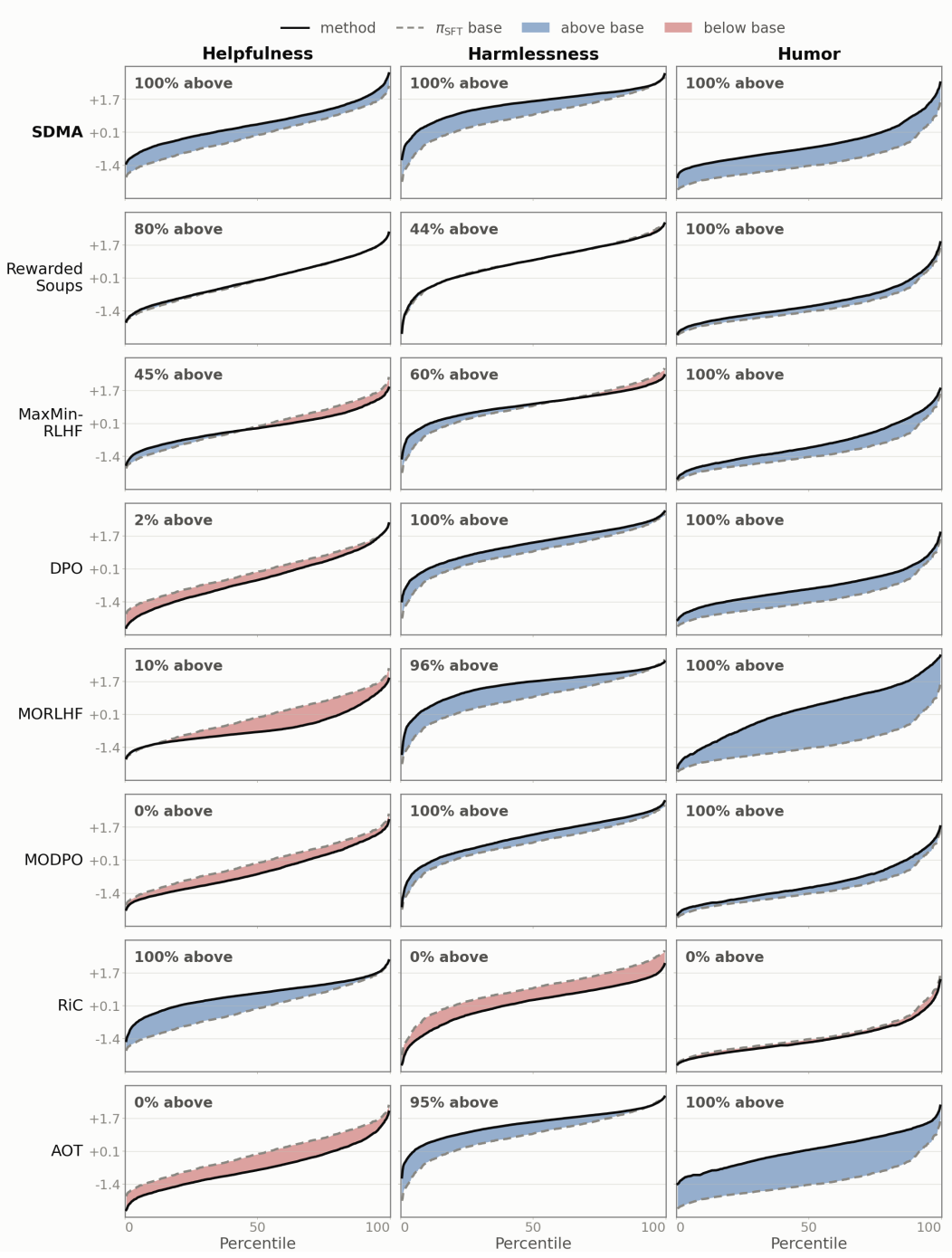


Figure 4: Per-objective reward quantile functions in the three-objective setting based on 5000 responses (1000 for SFT base) from 1000 prompts. Percentages indicate the fraction of plotted percentiles at which each method is at least as high as the reference. The y-axes are in standard deviations of the reference policy. Only SDMA is consistently above the base for each objective.

Table 5: Results with four objectives: helpfulness, harmlessness, humor, and honesty. The first five metrics are percentages. Effect sizes are Cohen's d relative to the reference policy. Best-of-16 rows are oracle references that use the evaluation reward models at inference time. Top values for non-oracle methods are shown in bold.

Method	MPD	$\mathrm{Lin}_{.05}$	$\mathrm{Lin}_{.10}$	$\mathrm{Lin}_{.05}^{\mathrm{WG}}$	$\mathrm{Lin}_{.10}^{\mathrm{WG}}$	d_{help}	d_{harm}	d_{humor}	d_{honesty}
SDMA	**98.1**	**100.0**	**98.5**	**99.9**	**76.1**	$+0.33$	$\mathbf{+0.72}$	$+1.32$	$+0.64$
Rewarded Soups	17.3	3.2	0.0	0.0	0.0	$+0.07$	-0.01	$+0.12$	$+0.29$
MODPO	8.2	0.6	0.0	0.0	0.0	-0.04	$+0.06$	$+0.12$	-0.10
DPO	19.8	93.7	73.9	0.5	0.0	-0.11	$+0.36$	$+0.98$	$\mathbf{+0.84}$
MaxMin-RLHF	10.9	42.9	2.3	4.7	0.0	-0.04	$+0.11$	$+0.41$	$+0.03$
MORLHF	9.9	85.4	61.8	0.7	0.0	-0.28	$+0.45$	$+0.99$	$+0.11$
AOT	0.3	50.8	40.0	0.0	0.0	-0.83	$+0.01$	$\mathbf{+2.47}$	-1.03
RiC	0.6	9.6	2.5	0.0	0.0	$\mathbf{+0.72}$	-0.65	-0.34	-0.33
Reference-policy	7.3	0.0	0.0	0.0	0.0	-0.01	$+0.00$	-0.01	$+0.01$
BoN-16 (π_{SFT}, oracle)	66.1	95.9	70.6	37.5	0.1	$+0.15$	$+0.18$	$+0.84$	$+0.73$
BoN-16 (SDMA, oracle)	99.5	100.0	99.8	100.0	95.1	$+0.51$	$+0.89$	$+1.62$	$+0.95$

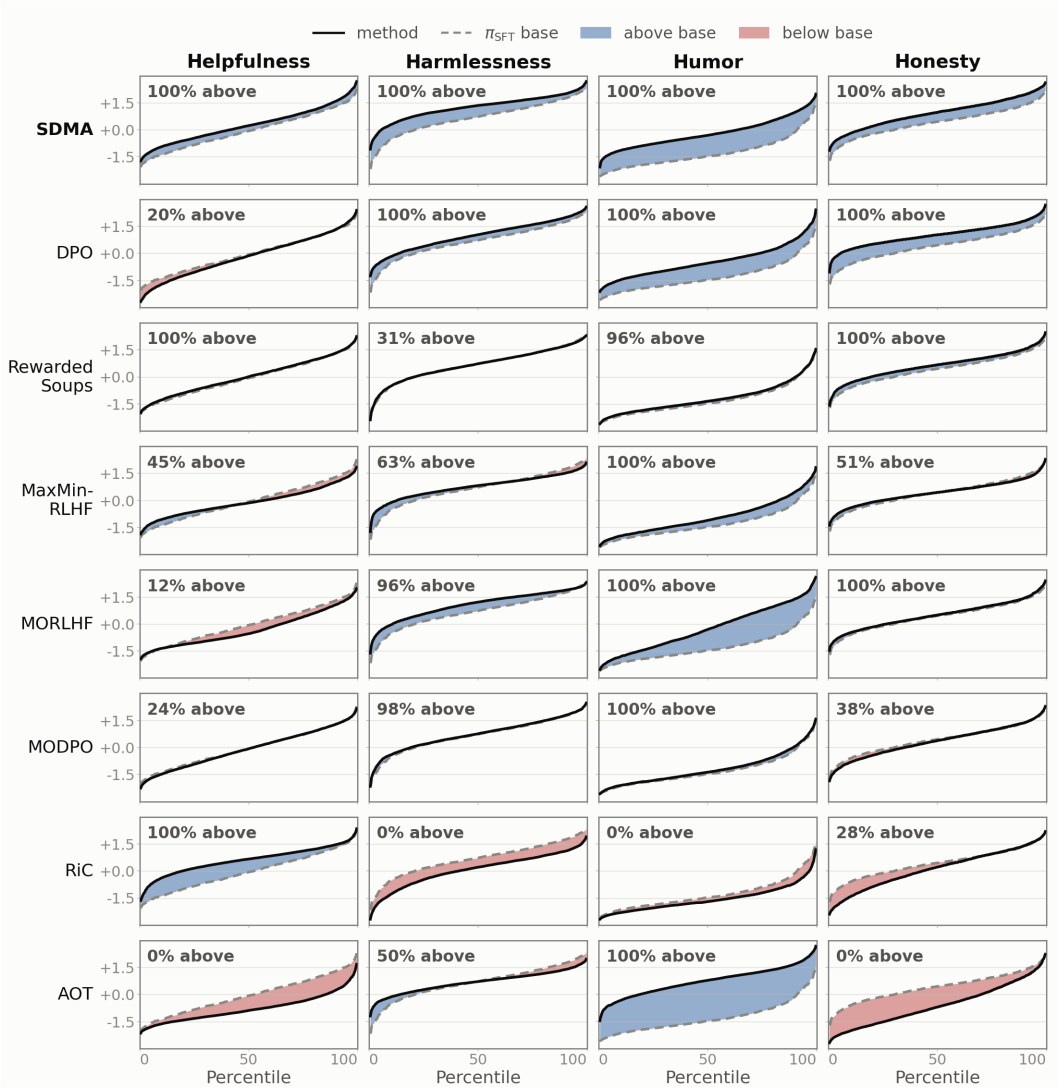


Figure 5: Per-objective reward quantile functions in the four-objective setting based on 5000 responses (1000 for SFT base) from 1000 prompts. Percentages indicate the fraction of plotted percentiles at which each method is at least as high as the reference. The y-axes are in standard deviations of the reference policy. Only SDMA is consistently above the base for each objective.